ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Econometrics and Management Science

Quantitative Finance

Selecting the prior shrinkage parameter in Bayesian VARs

Wessel Vis (409173)

Supervisor: dr. A.M. Camehl

Second Assessor: dr. J.W.N. Reuvers

April 28, 2021

Abstract

Bayesian vector autoregressions (BVARs) are an important tool for the modeling of the macroeconomy. However, for a long time these models could only contain a small number of variables, otherwise the number of parameters grew to be unwieldy and estimation became infeasible. This changed when De Mol et al. (2008) showed that shrinkage could be used to reliably estimate parameters even with a large number of variables in the BVAR. Several methods of selecting the amount of shrinkage have been developed that improve on past ad hoc heuristics by taking data-driven approaches. This paper investigates the merits of two promising methods. The first procedure aims to find the level of shrinkage that results in a target in-sample fit. The second uses a hierarchical prior and automatically selects the appropriate amount of shrinkage by maximizing the marginal likelihood. We find no significant differences in forecasting performance, short or long-term. Structural analysis shows that both methods can accurately capture macroeconomic dynamics through the impulse response functions.

Contents

1 Introduction **2**

2 Methodology **4**

2.1 Defining the priors 4

2.2 Choosing the amount of shrinkage 8

2.3 Forecasting with BVARs 9

2.4 Forecast evaluation 11

2.5 Impulse response functions 12

3 Data **13**

4 Forecasting with BVARs **14**

5 Robustness analysis **18**

6 Posterior distributions **21**

6.1 Posterior distribution of the hyperparameter 21

6.2 Posterior distribution of the VAR coefficients 22

7 Structural analysis **23**

8 Conclusion **27**

A Variables **30**

B Derivation of posterior parameters **32**

C Additional Results **34**

1 Introduction

In the field of macroeconomics the vector autoregressive (VAR) model is one of the most important tools available to the researcher. After Sims (1980) introduced VAR models to economics they have quickly been adopted for the use of macroeconomic forecasting and structural analysis. VAR models lend themselves well to capturing the linear interdependencies of multiple time-series. Without restrictions, the VAR model is very general and able to capture complex dynamics between the variables, but it has a lot of parameters that need to be estimated. Since the sample size that is typically available in macroeconomic applications is not enough to reliably estimate a large number of parameters, the researcher is faced with a dilemma: estimating an overparametrized system that will inevitably lead to unstable estimates, or limiting the number of variables, which will in turn result in an omitted variable bias. Past literature has proposed different solutions to this problem. One example is to use factor models to impose restrictions on the covariance structure, reducing the number of parameters to estimate (Forni et al., 2000; Stock & Watson, 2002). Other solutions often involve other types of restrictions, e.g. exclusion, exogeneity or homogeneity restrictions (Canova & Ciccarelli, 2004; Dees et al., 2007).

The seminal work of Bańbura, Giannone, and Reichlin (2010) shows that Bayesian shrinkage can be used to accurately estimate unrestricted VARs, even when they contain a large number of variables. They build on previous research from Litterman (1980, 1986) and others (Doan et al., 1984) that suggests that even for VARs of a modest size Bayesian shrinkage can improve the forecasting performance. The prior used in the paper by Bańbura et al. (2010) is based on the Minnesota prior (Litterman, 1979) but is adapted to include findings from the asymptotic analysis by De Mol, Giannone, and Reichlin (2008). This analysis shows that as the cross-sectional dimension of a Bayesian regression grows, the degree of shrinkage should be increased. In Bayesian vector autoregressions (BVARs) the amount of shrinkage is determined by a prior hyperparameter. The value of this hyperparameter is crucial for the performance of the model as it reflects the importance that the researcher puts on the prior beliefs. The early literature uses mostly ad hoc methods to set the value for this tightness parameter. Litterman (1980) maximizes the forecasting performance over a presample, Sims and Zha (1998) use fixed values for hyperparameters, and Bańbura et al. (2010) choose a value that achieves a desired in-sample fit.

A more inference-based approach to setting the hyperparameter is provided by Giannone, Lenza, and Primiceri (2015), they use a hierarchical modeling framework, where the procedure automatically selects the appropriate amount of shrinkage. Their findings align with the results in Bańbura et al. (2010) and De Mol et al. (2008), showing that larger BVARs benefit from more shrinkage. Our research analyzes the impact of different hyperparameter specifications on model performance in Bayesian vector autoregressions.

The main focus is on two different methods of estimating BVARs, that only differ in how the amount of shrinkage is chosen. The first is based on Bańbura et al. (2010), we use a natural conjugate prior that imposes the beliefs from the Minnesota prior. The shrinkage parameter is fixed to attain a certain level of in-sample fit. The second method resembles the hierarchical BVAR from Giannone et al. (2015), where a gamma distribution is set as a hyperprior for the tightness parameter. There are minor differences between the methods as they are described in the papers and our application of them. In Bańbura et al. (2010) an additional prior is set on the sum of coefficients and all prior beliefs are imposed by adding dummy observations. The hierarchical prior as presented in the second paper has a lot more than just one hyperparameter, the diagonal elements of the error covariance matrix and the tightness of two extra priors are all treated as hyperparameters. In our analysis the additional priors are omitted and we set prior densities on parameters, this ensures that the only difference across the methods is the way in which the shrinkage is chosen.

The hierarchical BVAR from Giannone et al. (2015) aims to set the shrinkage at a level that maximizes the marginal likelihood as a way of ensuring optimal forecasting performance at a one-step horizon. Putting this approach opposite the targeted fit, we might expect forecasts made by the hierarchical BVAR to perform better at forecasting short horizons. While the method from Bańbura et al. (2010) might lead to a more accurate modeling of the interdependencies of the variables, resulting in more plausible impulse response functions and more accurate forecasts at further horizons. Slight partial evidence for this hypothesis can be found in the results of Giannone et al. (2015). Their findings show that the hierarchical BVAR has a slight edge in performance at forecasting one step ahead, which is lost for some variables when forecasting four quarters ahead.

The analysis is set up such that the methods differ only in the way shrinkage is selected. The two different procedures, along with a fixed shrinkage BVAR, are evaluated by estimating VAR models that differ in size. These models are constructed with variables from a large set of U.S. macroeconomic time series, which contains monthly observations over the period 1960-2003. The estimated models are compared based on their ability to predict key variables at different horizons, up to one year ahead. We look at point forecasts as well as predictive densities. As a structural analysis, we also investigate the effect of a monetary policy shock on the system by estimating the impulse response functions and seeing if they are in line with expectations from literature and theory.

We find that all BVARs can beat the benchmark model of a random walk with drift at short horizons, but the difference in performance between the BVARs is mostly minimal. One of the causes for this is that the methods arrive at similar shrinkage values. We also see that shrinkage can benefit the forecasting performance in VARs with as few as three variables. Furthermore, we confirm two observations from Bańbura et al. (2010) on larger

BVARs: as the size of a BVAR increases, 1) the priors should be set tighter to prevent over-fitting 2) the estimated size of the non-systematic component of monetary policy decreases.

The remainder of the paper is structured as follows. section 2 explains the priors and evaluation measures in more detail. section 3 provides more information on the dataset and defines the VAR models. In section 4 and section 5 we compare the forecasting performance and conduct a rudimentary robustness analysis. section 6 investigates the posterior distributions of the parameters, followed by the structural analysis in section 7. section 8 concludes.

2 Methodology

Vector autoregressive models offer a very general representation and can capture complex data relationships, but this brings with it the risk of overparametrization. Given the usual sample size in macroeconomic applications, only a small number of free parameters can be confidently estimated. Because of this, traditionally, the number of variables that can be included is very limited. Typically, three to ten variables are used when estimating VARs. This, in turn, introduces a high potential for omitted variable bias. Bańbura et al. (2010) show that Bayesian shrinkage allows the handling of large unrestricted VARs. Motivated by results in De Mol et al. (2008), they increase the tightness of the priors as the models become larger.

2.1 Defining the priors

Let us first consider a VAR model of order p for a vector y_t containing n dependent variables,

$$y_t = c + B_1 y_{t-1} + \ldots + B_p y_{t-p} + u_t \quad \text{for } t = 1, \ldots, T, \tag{1}$$

where the n-dimensional residuals u_t have a normal distribution with zero mean and covariance matrix $\text{Var}(u_t) = \Sigma$, c is an $n \times 1$ vector of intercepts and $B_1, \ldots, B_p$ are $n \times n$ parameter matrices.

It is clear that VAR models are far from parsimonious, there are n equations, each of which contains $1+np$ parameters. This means that a model for 5 dependent variables that considers 4 lags will already have more than 100 coefficients to estimate. With the limited sample size that is usually available in macroeconomic applications it becomes hard to obtain precise estimates, making it necessary to introduce extra prior information into the system. The applications in this paper use the Bayesian method of imposing prior densities on the parameters.

Both the method of Bańbura et al. (2010) and Giannone et al. (2015) use the Min-

nesota prior as a starting point, with adjustments proposed by Kadiyala and Karlsson (1997) and Sims and Zha (1998). The Minnesota prior was put forward by Robert Litterman in the early 80's and quickly became popular for its simplicity and effectiveness. In Litterman (1979) he observes that it is common for economic variables to have a random walk component and proposes to center the equations around a random walk with drift. This is accomplished by setting the prior mean for the diagonal elements of B_1 to 1, for the remaining coefficients the mean is set at 0. Litterman (1986) argues that it is reasonable to assume that more recent lags hold more pertinent information in explaining future values, this belief is incorporated by shrinking the coefficients on lags of a higher order more towards 0. These beliefs can be represented by the following moment definitions:

$$\text{Var}[(B_r)_{ij}] = \frac{\lambda^2}{r^2}\frac{\sigma_i^2}{\sigma_j^2}, \tag{2}$$

where i and j are variables indices, and r is the lag length. The coefficients on lags of different variables are not scale invariant. To account for the differing levels of variability between the variables, the term σ_i^2/σ_j^2 is included. Here σ_i is a measure of the variance of variable i, for which Litterman (1986) uses s_i, the estimated standard error of the residuals of an AR(1) model on variable i. Note that for lags of variables in their own equation, $i = j$, so the term σ_i^2/σ_j^2 disappears and the variance of coefficients on the diagonal of B_r is equal to λ^2/r^2.

The hyperparameter λ functions as a tightness parameter that reflects the weight that the researcher puts on the prior beliefs. For example, if $\lambda = 0$, the prior means become dogmatic and the data has no impact on posterior results. But if we set $\lambda = \infty$, estimates are completely determined by the data and are equal to the estimates obtained from an ordinary least squares regression. Bańbura et al. (2010) argue that the researcher should let λ depend on the size of the equation. A larger number of endogenous variables should correspond to a tighter prior that shrinks more to prevent over-fitting (this is also shown in De Mol et al., 2008). We let prior variances decrease quadratically with lag order $(1/r^2)$. Early implementations (Kadiyala & Karlsson, 1997; Litterman, 1979) of the Minnesota prior would sometimes use a linear rate $(1/r)$, but we take the more common view that the relevance of past observations drops off more quickly.

The specification of the Minnesota prior is completed by choosing a normal distribution for the coefficients, a diffuse prior on the intercept, and setting the covariance matrix of the residuals Σ fixed and diagonal, with diagonal elements $\Sigma_{ii} = s_i^2$.

This last condition especially is considered problematic as it is difficult to justify the assumption of uncorrelated errors in a macroeconomic setting. We therefore follow Kadiyala and Karlsson (1997) in their use of an inverted Wishart distribution as a prior for Σ. This combination of distributions is known as the Normal-inverse-Wishart prior and it is the

natural conjugate prior for data that is assumed normal. To implement these priors it is preferable to use a different representation where the VAR(p) model is stacked over t. The dependent variables are stacked into a $T \times n$ matrix $Y = (y_1, \ldots, y_T)'$, so that row i is y_i'. The regressors are grouped together in $X_t = (1, y_{t-1}', \ldots, y_{t-p}')'$, to create vectors of length $k = np + 1$, which are then stacked to form the $T \times k$ matrix $X = (X_1, \ldots, X_T)'$. We gather the coefficients together in $B = (c, B_1, \ldots, B_p)'$ with dimensions $k \times n$, so that (1) can be written as

$$Y = XB + U, \tag{3}$$

where $U = (u_1, \ldots, u_T)'$ is the $T \times n$ matrix of innovations.

The normal-inverse-Wishart prior is then defined on these variables by a matricvariate normal distribution on the coefficient matrix B and an inverted Wishart distribution on the covariance matrix of the disturbances Σ

$$B|\Sigma \sim MN(\beta_0, \Sigma \otimes \Omega_0) \quad \text{and} \quad \Sigma \sim IW(S_0, d_0). \tag{4}$$

Where $MN(\cdot)$ is the matricvariate normal distribution, to reflect the beliefs described earlier, its mean β_0 is a $k \times n$ matrix[1] filled mainly with zeros except for the elements corresponding to the diagonal of B_1. The scale matrix of the inverse Wishart distribution, S_0, is diagonal with elements s_i^2. The prior mean of Σ is equal to $S_0/(d_0 - n - 1)$, so for the prior degrees of freedom we choose $n + 2$, giving a prior mean of S_0. The moments from (2) are then maintained by setting $\Omega_0 = \text{diag}(w)$ and constructing w in the following way

$$w_q = \begin{cases} \frac{\lambda^2}{r^2} \frac{1}{s_j^2} & \text{if} \quad q \neq 1 \\ \kappa & \text{if} \quad q = 1 \end{cases} \tag{5}$$

where $r = \lfloor \frac{q-2}{n} \rfloor + 1$ and $j = q - 1 - \lfloor \frac{q-2}{n} \rfloor n$. The first element of the diagonal corresponds to the intercepts and is set to a large number κ to establish an uninformative prior. Following Kadiyala and Karlsson (1997) we use the value $\kappa = 10^7$. With this definition of Ω the Kronecker product of Σ and Ω reproduces the prior covariances as they were previously established.

Because this combination of priors is a natural conjugate, the posterior distribution is

[1] This notation deviates from the convention of using β for the vectorized form of the VAR coefficients B, instead β_0 and β_1 denote the prior and posterior means of B and they therefore have the same dimensions as B.

of the same form, a conditional matricvariate normal and an inverse Wishart distribution

$$
\begin{aligned}
& B|\Sigma \sim MN(\beta_1, \Sigma \otimes \Omega_1) \quad \text{and} \quad \Sigma \sim IW(S_1, d_1), \\
\text{where} \quad & \beta_1 = \left(X'X + \Omega_0^{-1}\right)^{-1} \left(X'Y + \Omega_0^{-1}\beta_0\right), \\
& \Omega_1 = \left(X'X + \Omega_0^{-1}\right)^{-1}, \\
& S_1 = S_0 + (Y - X\beta_1)'(Y - X\beta_1) + (\beta_1 - \beta_0)' \, \Omega_0^{-1} \, (\beta_1 - \beta_0), \\
& d_1 = d_0 + T.
\end{aligned}
$$

The derivations of the posterior parameters can be found in Appendix B. From these definitions it is clear that the posterior variance of the coefficients is inversely proportional to the term $X'X$, which is a measure of the variability in the data, as well as proportional to the prior variance. While the posterior mean is a weighted average of the prior mean and the OLS estimate. To elucidate the role of λ, consider decomposing Ω_0 into two terms, the shrinkage factor and a covariance matrix that adheres to the beliefs set out previously, $\Omega_0 = \lambda^2\Omega^*$. The matrix Ω^* is constructed in a way very similar to (5), but with elements $\frac{1}{r^2 s_j^2}$, so this matrix represents the prior belief that coefficients on higher lags should be shrunk more towards zero. The formulae of the posterior mean and variance of the coefficients can then be written as

$$
\begin{aligned}
\Omega_1 &= \left(X'X + \frac{1}{\lambda^2}\Omega^{*-1}\right)^{-1} \\
\beta_1 &= \left(X'X + \frac{1}{\lambda^2}\Omega^{*-1}\right)^{-1} \left(X'Y + \frac{1}{\lambda^2}\Omega^{*-1}\beta_0\right).
\end{aligned}
$$

From this formulation one can more clearly see the influence of λ on the posterior parameters. The prior information enters the equations through β_0 and Ω^*, the shrinkage term regulates their weight relative to the information from the observations in the data. As the hyperparameter is set to smaller values and gets closer to zero, the influence of the prior becomes more dogmatic until $\lim_{\lambda \to 0} \beta_1 = \beta_0$. On the other hand, setting λ at increasingly larger values diminishes the contribution of the prior information in the posterior parameters, $\lim_{\lambda \to \infty} \beta_1 = (X'X)^{-1}X'Y = \hat{B}_{\text{OLS}}$.

In Bańbura et al. (2010) the informativeness of the priors is set at a value that achieves a desired level of in-sample fit. The authors of Giannone et al. (2015) take a different approach and make the argument that the tightness of the priors should be treated as another parameter. They suggest the use of a hierarchical modeling procedure, where a prior is put on the hyperparameters as well. This method would make it possible to perform inference on the hyperparameters, giving further insight into the role of prior distributions. To do so, we would need their posterior. If we let β^+ be the vector that

collects all parameters from the VAR model, and δ is a vector of hyperparameters with corresponding hyperprior $p(\delta)$, then the posterior of δ is given by

$$p(\delta|y) \propto p(y|\delta)p(\delta). \tag{6}$$

Here $p(y|\delta)$ is the marginal likelihood (ML) of the data, which can be obtained by integrating the parameters β^+ out of the joint likelihood

$$p(y|\delta) = \int p(y|\beta^+, \delta)p(\beta^+|\delta)d\beta^+.$$

From (6) it is clear that when a flat hyperprior is used, the posterior is proportional to this marginal likelihood. It can be shown that maximizing the posterior likelihood of the hyperparameters then becomes equivalent to maximizing the predictive densities of one-step-ahead forecasts

$$p(y|\delta) = \prod_{t=1}^{T} p\left(y_t|y^{t-1}, \delta\right).$$

The ML is rewritten into the product of conditional densities of the individual observations. Combined with (6) this shows that, under a flat hyperprior, maximizing the ML equates to optimizing the predictive capability of the model at the one-step horizon.

2.2 Choosing the amount of shrinkage

In this paper we will focus on two different approaches to setting the amount of shrinkage. The first focuses on attaining a certain level of in-sample fit. We use the same measure for fit that is defined in Bańbura et al. (2010). Before we explain this method, we first define the mean squared forecast error (MSFE)

$$\text{MSFE} = \sum_{t=1}^{T} \left(\hat{y}_t - y_t\right)^2.$$

This metric takes the average of the squared forecast errors over a certain sample, when measuring forecast performance this can be the sample of a pseudo-out-of-sample forecasting exercise. In this particular instance, we use it as a measure of in-sample fit, meaning the $\hat{y}_t$ are the fitted values in a pre-sample. The specific metric that is used in Bańbura et al. (2010) is the average of the ratios of the MSFEs obtained by estimating a simple model and a random walk in a pre-evaluation period. We set the target fit at the level that is obtained by estimating a small benchmark model with $\lambda = \infty$, so that it becomes equivalent to OLS. The target fit is defined as

$$\text{Fit} = \frac{1}{3} \sum_{i \in \mathcal{I}} \frac{\text{msfe}_i^{\infty}}{\text{msfe}_i^{0}},$$

where msfe_i^{λ} is the MSFE score that is achieved for variable i over the presample with shrinkage set at λ, and $\mathcal{I}$ is defined as the set that contains the three key variables in the simple model. The uncapitalized letters indicate that this is an MSFE score over the pre-evaluation period. When we have found the desired level of fit, we then use a grid search to find the value for λ that best achieves that degree of fit

$$\lambda = \arg\min_{\lambda} \left| \text{Fit} - \frac{1}{3} \sum_{i \in \mathcal{I}} \frac{\text{msfe}_i^{\lambda}}{\text{msfe}_i^{0}} \right|.$$

This value is then used to estimated the BVAR at each iteration throughout the sample. Note that when we use this method to find a shrinkage value for the small benchmark model, it will always select $\lambda = \infty$, i.e. no shrinkage.

For the hierarchical BVAR, we do not directly set a value for λ, but we impose a hyperprior.

$$p(\beta, \Sigma, \lambda | y) \propto p(\beta | \Sigma, \lambda) p(\Sigma | \lambda) p(\lambda) p(y | \beta, \Sigma, \lambda)$$

In this case inference on λ can be performed by deriving the marginal posterior. This is done by integrating the model parameters out of the joint posterior

$$p(\lambda | y) = \int p(\beta, \Sigma, \lambda | y) \, d\beta \, d\Sigma$$

Following Giannone et al. (2015) we use a Gamma distribution as a hyperprior. For the parameters of this distribution we use values that result in the characteristics recommended by Sims and Zha (1998), a mode of 0.2 and a standard deviation equal to 0.4.

2.3 Forecasting with BVARs

After the models have been estimated there are a two ways in which we create forecasts, the first is to use the posterior mode of the parameter estimates to compute point forecasts. The second method reflects the uncertainty that the Bayesian models put on the parameter estimates by producing a predictive density. This density is not available in its unconditional form, so it is approximated by draws that are made using the posterior distributions. These random draws can then be used for inference.

The posterior mode of the coefficient matrices corresponds to the value that the model has deemed the most likely after combining the information in the priors and the data. They are used to create point forecasts of future observations

$$\hat{y}_{T+1|T} = \hat{c} + \hat{B}_1 y_T + \ldots + \hat{B}_p y_{T-p+1}, \quad \text{for } T = T_{\text{start}}, \ldots, T_{\text{end}}.$$

Where the hat superscript is used to denote a model estimate and the $|T$ subscript signifies that the forecast is made at time T, i.e. only information available at that point is used.

T_{start} and T_{end} are the first and last observation of the sample used for the pseudo-out-of-sample forecasting exercise.

Forecasts at horizons further ahead than one step are computed iteratively as follows

$$\hat{y}_{t+h|t} = \hat{c} + \hat{B}_1 \hat{y}_{t+h-1|t} + \ldots + \hat{B}_p \hat{y}_{t+h-p|t}, \quad \text{for } t = T_{\text{start}}, \ldots, T_{\text{end}}, \tag{7}$$

where $\hat{y}_{t|T}$ is defined as the observed value y_t if $t \leq T$, or the model estimate otherwise. As h increases, observed values get replaced by model predictions of future values.

The predictive densities cannot be analytically evaluated, only conditional posteriors are available in their analytical form. We can still perform inference on the predictive likelihoods by drawing from them using two different sampling algorithms. When we estimate the BVAR with a fixed hyperparameter the posterior density is of a convenient form, because of the natural conjugate prior. In this case draws can be sampled from the conditional distribution.

1. Get a draw $\Sigma^{(i)}$ from its posterior distribution $IW(S_1, d_1)$.
2. Get a draw $B^{(i)}$ from the conditional posterior distribution $MN(\beta_1, \Sigma^{(i)} \otimes \Omega_1)$.
3. Draw an innovation from a Gaussian noise density: $u_t^{(i)} \sim N(0, \Sigma^{(i)})$.
4. Sample from the predictive density by multiplying with the predictive variables and adding the innovation draw $y_{T+1|T}^{(i)} = X_T B^{(i)} + u_T^{(i)}$.

Repeating these steps N times results in a set $\{y_{T+1|T}^{(i)}\}_{i=1}^{N}$, which can be used as an empirical predictive likelihood. To create forecasts at horizons further than one, use (7) in the last step, with $B^{(i)} = (\hat{c}, \hat{B}_1, \ldots, \hat{B}_p)'$. The series of draws for B and Σ are later used in the calculation of the impulse response functions. Sampling in the case of the hierarchical prior is slightly more complicated, since we cannot directly make draws for the hyperparameter we have to resort to some MCMC algorithm. Giannone et al. (2015) suggest the following Metropolis-Hastings sampler:

1. Initialize $\lambda^{(0)}$ at its posterior mode, obtained using numerical maximization. Set iteration counter, $i = 1$.
2. Draw a candidate value λ^* from the proposal distribution $N(\lambda^{(i-1)}, cH^{-1})$, where H is the Hessian of the negative of the log-posterior of λ evaluated at the mode, and c is a scaling constant.
3. Compute the acceptance probability $\alpha^{(i)} = \min\left\{1, \frac{p(\lambda^*|y)}{p(\lambda^{(i-1)}|y)}\right\}$. Accept the candidate draw (i.e. $\lambda^{(i)} = \lambda^*$) with probability $\alpha^{(i)}$, otherwise set $\lambda^{(i)} = \lambda^{(i-1)}$.

4. Draw $\Sigma^{(i)}$, followed by $B^{(i)}$ and $u_t^{(i)}$ and compute $y_{t+1|t}^{(i)}$ using the previous sampling scheme.

5. Set $i = i + 1$, go to 2.

The posterior mode of λ is found by maximization of the marginal likelihood, this is done in MATLAB with Christopher Sims' function `csminwel.m`[2], which uses a quasi-Newton method with BFGS updates of the inverse Hessian. It is designed to be robust against certain characteristics of likelihood functions that can make optimization difficult. The final approximation of the inverse Hessian from this function is used for the proposal distribution. The value of the scaling constant c should be chosen to achieve a desired acceptance rate. It is shown in Gelman et al. (1996) that the optimal acceptance rate for a univariate target distribution is approximately 44%.

2.4 Forecast evaluation

Evaluation of the point forecasts is performed using the ratios of MSFEs. For the purpose of this forecasting exercise we provide a more elaborate definition of the MSFE than the one previously given

$$\text{MSFE}_{i,h}^{M} = \frac{1}{T_{\text{end}} - T_{\text{start}} + 1} \sum_{t=T_{\text{start}}}^{T_{\text{end}}} \left(\hat{y}_{i,t+h|t}^{M} - y_{i,t+h} \right)^2,$$

where M is the model used to forecast, i is a specific variable contained in the model, and h is the forecast horizon. The subscript $|t$ is used to reflect that only information up to time t is used in making the forecast. Results are reported relative to a benchmark MSFE:

$$\text{RMSFE}_{i,h}^{M} = \frac{\text{MSFE}_{i,h}^{M}}{\text{MSFE}_{i,h}^{0}}.$$

Where $\text{MSFE}_{i,h}^{0}$ is the value of the metric obtained by the benchmark for variable i at horizon h. A value of one for this statistic indicates a forecasting performance on par with the benchmark, values below one mean that the BVARs perform better.

A metric that is better suited for comparing Bayesian models is the log predictive score, because it takes into account the uncertainty of the predictions. In Giannone et al. (2015) the log predictive score is measured using Gaussian approximations for all predictive densities. To compare the performance of two competing models, we take the average difference of the scores.

$$\text{DLPS}_{M_1,M_2} = \frac{1}{T_{\text{end}} - T_{\text{start}} + 1} \sum_{t=T_{\text{start}}}^{T_{\text{end}}} \log \frac{p_{M_1}(y_t|y)}{p_{M_2}(y_t|y)},$$

[2]The function and related files are available for download at Christopher Sims' personal page.

where $p_{M_i}(y_t|y)$ is the predictive density produced by model M_i evaluated at the realized value y_t.

2.5 Impulse response functions

The estimated VAR models provide us with representations of the macroeconomic variables and their interdependencies. Because of this they can be used to identify structural shocks and assess their transmission mechanism. One example of such an application where this is useful is when policy makers want to investigate the wider effect of a proposed change in monetary policy. This is achieved through a structural analysis using the impulse response functions (IRFs), these enable us to calculate the effects of an exogenous shock. Before this can be done, the VAR needs to be rewritten into its structural form. The goal is to find the effects of a monetary policy shock, so we need to isolate these effects, which cannot be done with the VAR in (1). Since the assumption of independent shocks is not reasonable in a macroeconomic setting, the VAR in (1) is structured to have non-zero dependence, which means a single shock cannot be applied. To remedy this, consider the following. Let P be the lower diagonal Cholesky matrix of the covariance of the disturbances, i.e. $\mathrm{E}(u_t u_t') = \Sigma = PP'$. Then $e_t = P^{-1}u_t$ is defined as a linear transformation of the VAR innovations. Apply the same transformation to the rest of the VAR model in (1) to get

$$\mathcal{B}_0 y_t = \nu + \mathcal{B}_1 y_{t-1} + \ldots + \mathcal{B}_p y_{t-p} + e_t \quad \text{with} \quad e_t \sim N(0, I), \tag{8}$$

where $\nu = P^{-1}c$, $\mathcal{B}_0 = P^{-1}$, and $\mathcal{B}_i = P^{-1}B_i$ for $i = 1, \ldots, p$. The covariance matrix of the disturbances e_t is diagonal, so a singular shock can be entered into the system. Note that since P is lower triangular, depending on the order of the variables, some variables are affected contemporaneously. Therefore, the monetary policy shock is identified using a standard recursive identification scheme that takes this into account. The scheme works by dividing the variables in the dataset based on how fast they react to an unforeseen change in monetary policy. Following Christiano et al. (1999), Bernanke et al. (2005), Stock and Watson (2005b), we use two categories: fast-moving and slow-moving. The variables are then ordered as $y_t = (x_t, r_t, z_t)$. Here x_t contains the n_1 slow-moving variables that are assumed to not react contemporaneously to the shock in monetary policy, such as prices and real variables, r_t is the monetary policy instrument, and the fast-moving variables are gathered in z_t, these are expected to be affected contemporaneously by the shock. Mostly, z_t are financial variables. In the experiment, the shock is then applied to element $n_1 + 1$ of e_t.

With the appropriate order of variables the IRFs can be correctly computed, similarly to Canova (1991) and Gordon and Leeper (1994). For each draw (B, Σ) from the posterior, we calculate P and $\mathcal{B}_i$ for $i = 0, \ldots, p$, followed by calculating the effects of the shock on

the variables

$$g_{t+h} = \nu + \mathcal{B}_1 g_{t+h-1} + \ldots + \mathcal{B}_p g_{t+h-p} + \mathbf{1}_{h=1} e_{t+h}$$

Here g_t is used to denote the impact of the shock on the variables, $\mathbf{1}_{h=1}$ is an indicator function that takes the value one when $h = 1$ and zero otherwise. The shock enters through innovation e_{t+1} which has one non-zero element in position $n_1 + 1$.

3 Data

We use the macroeconomic data set from Stock and Watson (2005a), which contains monthly observations on 131 US macro indicators. The time span covered by this data set is January 1960 to December 2003. It contains 540 observations on the included variables. The methods will be tested by using them on models of different sizes. The first model contains three key variables (Small), a measure of real economic activity (employees on nonfarm payrolls) a measure of the price level (core consumer price index), and a measure of monetary policy (federal funds rate). The second model will be medium-size at 7 variables (Medium), in addition to the three key variables it contains the index of sensitive material prices and three monetary variables: the total reserves and the non-borrowed reserves of depository institutions, and the M2 money stock. The third model consists of 20 variables (Large), it is created by expanding the Medium model with other macroeconomic variables that enrich the system with information on total consumption and production, the labor market, the housing market, and the financial markets.

A few of the variables are already expressed in annual rates, these we take in levels. To the other variables we apply a logarithmic transformation. As discussed in section 2, the prior means on the coefficients are set to reflect the belief that the random walk with drift is an accurate approximation of the behavior of the variable. This is true for variables with a high degree of persistence, but not for variables that are better characterized by a mean-reverting process. We therefore perform a few statistical tests to determine whether the random walk prior is correct for every variable in our dataset. We conduct the augmented Dickey-Fuller test and the Phillips-Perron test that both test for the presence of a unit root, as well as the Kwiatkowski-Phillips-Schmidt-Shin test with a null hypothesis of stationarity. The results are in Table 6. Conclusive evidence of stationarity is only shown for the variable housing starts (HOUST), so when setting the prior mean on the first coefficient matrix B_1, the diagonal element corresponding to this variable is set to 0, while the other diagonal entries are 1. Table 5 in Appendix A presents the dataset concisely, it contains all information discussed in this section as well as the categorization into fast-moving and slow-moving variables for the purpose of the strucural analysis.

The first moment in time that we start estimating and making forecasts is June 1971 (T_{start}), at each point from then up until December 2002 (T_{end}), we use the previous 120

observations to estimate the models and make forecasts for four horizons, up to 12 months ahead. At each of the 379 observations, the three BVARs of different sizes are estimated with three methods of determining the amount of shrinkage, resulting in nine estimated BVARs at every one of those points. Each of these models is then used to make point forecasts and produce predictive densities for the four different horizons.

4 Forecasting with BVARs

We start off with the evaluation of point forecasts made using the different BVARs. These methods only differ in how the amount of shrinkage is chosen. The first is based on Bańbura et al. (2010), here we set λ at the value that best achieves the fit in a presample of 10 years, from now on this is referred to as BGR. The second method follows Giannone et al. (2015), at each iteration we maximize the posterior distribution of λ, this peak value is then used for the point forecast, we call this GLP. For the third method we fix the value of λ at 0.2, the value suggested in Sims and Zha (1998), from now on referred to as Sims.

Table 1: Relative MSFE of point forecasts.

		Small			Medium			Large		
Horizon	Variable	BGR	GLP	Sims	BGR	GLP	Sims	BGR	GLP	Sims
$h = 1$	EMPL	0.60*	0.54*	**0.52***	0.60*	**0.60***	0.62*	**0.51***	0.52*	0.57*
	CPI	0.63*	**0.53***	0.53*	**0.52***	0.52*	0.54*	**0.46***	0.50*	0.56*
	FED	1.09	**0.96**	0.97	0.88*	0.89*	**0.86***	0.86*	**0.81***	0.85
$h = 3$	EMPL	0.46*	0.46*	**0.44***	0.56*	**0.55***	0.57*	**0.45***	**0.45***	0.51*
	CPI	0.56*	0.51*	**0.50***	**0.45***	0.46*	0.48*	**0.34***	0.41*	0.46*
	FED	1.31**	1.17	**1.15**	0.95	0.96	**0.94**	1.01	**0.97**	1.02
$h = 6$	EMPL	**0.52***	0.57	0.55*	**0.64**	0.64	0.67	**0.56**	0.60	0.68
	CPI	0.63	0.55	**0.54***	**0.46***	0.46*	0.49*	**0.33***	0.38*	0.40*
	FED	1.39	1.29	**1.26**	1.03	1.03	**1.03**	1.28	**1.26**	1.38
$h = 12$	EMPL	**0.60**	0.74	0.71	**0.74**	0.76	0.78	**0.74**	0.83	1.02
	CPI	0.70	0.60	**0.59**	0.53*	**0.53***	0.57	**0.37***	0.38*	0.38*
	FED	1.46	1.43	**1.38**	1.28	1.27	**1.25**	**1.51**	1.52	1.88**

Reported values are relative to the MSFE of a random walk forecast. Results are reported for the variables employment (EMPL), consumer price index (CPI) and federal funds rate (FED) at forecast horizons of one month, one quarter, six months, and one year ahead. The methods differ in the way λ is chosen: optimize the fit in a presample (BGR), maximize the posterior likelihood (GLP), or fixed value $\lambda = 0.2$ (Sims). The results are obtained using different model sizes: Small: 3 variables, Medium: 7 variables, Large: 20 variables. All BVARs are estimated with five lags ($p = 5$). For each combination of model, variable, and horizon the lowest score is printed in boldface. Values below one indicate that the methods outperform the benchmark. Asterisks indicate whether the difference in predictive ability is significant based on the Diebold-Mariano test.
*: Null hypothesis of equal predictive accuracy is rejected in favour of the BVAR at a confidence level of 5%.
**: The null hypothesis is rejected in favour of the benchmark at a confidence level of 5%.

The results in Table 1 show the forecasting performance of the BVARs relative to a random walk with drift. The presented values are ratios of the MSFE statistics of the

BVARs and the benchmark. We see that all three methods outperform the benchmark at almost every instance. However, the BVARs seem unable to better forecast the federal funds rate (FED) at horizons beyond one quarter. None of the three shrinkage methods consistently outperforms the other two, in a lot of the cases the methods lie very close together. This is especially true for GLP and Sims in combination with the Small and Medium model, the reason for this is that the hierarchical model uses shrinkage values between 0.2 and 0.35, while for Sims the value is fixed at 0.2. Litterman (1986) has shown that the forecasting performance of a BVAR exhibits little sensitivity to changes in shrinkage within this range. Another observation we can make is that forecasting performance for the CPI increases with model size at all horizons, indicating that the added variables in the larger models contain useful information in predicting the price level. We run all our VAR models with five lags, in section 5 we show that results are largely robust to changes in the lag order.

The asterisks in Table 1 mark where the difference in forecasting performance is significant. The significance is based on a two-sided Diebold-Mariano test with quadratic loss, the statistics themselves are provided in Appendix C. At the shorter horizons the BVARs dominate the benchmark in predicting employment and price level. For longer horizons this difference remains significant only for the price level. Lastly, we note the two instances where a BVAR is found to be significantly worse than the benchmark. Both occur for the FED, once at the three-month horizon (Small BGR) and once at the one-year horizon (Large Sims).

Comparing the point forecasts can be very useful in determining the relative performance of models. However, it does not capitalize on one of the great benefits of Bayesian models, the fact that each model provides a measure of the uncertainty of the forecast. The log predictive score is a simple measure that allows us to evaluate and compare this uncertainty. The predictive score is defined as the likelihood that a model puts on the realized value of a variable, which we compute by evaluating the predictive density at the observed value. Since the analytical expressions of unconditional predictive densities are not available, we use Gaussian approximations. This approximation is both simple and a close representation of the actual density. After taking the logarithms of these scores, we take the differences between two models to compare them against each other, the averages of these differences are reported in Table 2. The three different shrinkage methods are compared against the random walk with drift. A higher log predictive score means that the model assigned a higher likelihood to the realization of an observed value, therefore a positive statistic in Table 2 indicates that on average the BVAR outperformed the random walk prior.

The table also shows HAC estimates of the standard errors to give an indication of the statistical significance of the results. Looking at the results it is hard to find any

Table 2: Average difference of log predictive scores.

		Small			Medium			Large		
Horizon	Variable	BGR	GLP	Sims	BGR	GLP	Sims	BGR	GLP	Sims
$h = 1$	EMPL	0.31	0.29	0.29	0.26	0.24	0.27	0.32	0.35	0.33
		(0.71)	(0.65)	(0.63)	(0.70)	(0.73)	(0.70)	(0.66)	(0.68)	(0.75)
	CPI	0.29	0.36	0.36	0.33	0.33	0.33	0.34	0.31	0.27
		(1.03)	(0.88)	(0.89)	(0.87)	(0.88)	(0.87)	(0.86)	(0.84)	(0.90)
	FED	0.12	0.10	0.11	0.11	0.10	0.14	0.14	0.16	0.15
		(0.42)	(0.36)	(0.33)	(0.33)	(0.34)	(0.40)	(0.35)	(0.47)	(0.62)
$h = 3$	EMPL	0.43	0.35	0.35	0.26	0.22	0.29	0.38	0.37	0.34
		(1.44)	(1.46)	(1.42)	(1.58)	(1.53)	(1.61)	(1.52)	(1.52)	(1.58)
	CPI	0.61	0.65	0.65	0.64	0.63	0.57	0.71	0.60	0.57
		(2.21)	(2.05)	(2.04)	(1.85)	(1.86)	(1.86)	(1.81)	(1.82)	(1.79)
	FED	−0.11	−0.04	−0.02	0.06	0.05	0.04	0.04	0.07	0.04
		(0.99)	(0.95)	(0.94)	(0.85)	(0.83)	(0.97)	(0.81)	(0.90)	(1.05)
$h = 6$	EMPL	0.67	0.55	0.51	0.45	0.38	0.48	0.55	0.61	0.57
		(2.41)	(2.64)	(2.71)	(2.56)	(2.47)	(2.54)	(2.58)	(2.66)	(2.69)
	CPI	1.13	1.27	1.28	1.20	1.18	1.10	1.36	1.26	1.21
		(3.83)	(3.51)	(3.47)	(3.05)	(3.07)	(3.17)	(2.81)	(2.68)	(2.71)
	FED	−0.16	−0.13	−0.16	0.01	0.00	−0.01	0.01	0.04	0.00
		(1.48)	(1.66)	(1.67)	(1.28)	(1.26)	(1.34)	(1.38)	(1.42)	(1.41)
$h = 12$	EMPL	1.47	1.19	1.14	1.20	1.16	1.22	1.33	1.36	1.30
		(4.30)	(4.78)	(4.56)	(4.40)	(4.36)	(4.29)	(4.39)	(4.64)	(4.66)
	CPI	3.29	3.75	3.77	3.33	3.32	3.17	3.60	3.55	3.51
		(8.81)	(9.00)	(8.90)	(7.00)	(7.00)	(6.67)	(6.60)	(6.59)	(6.62)
	FED	0.00	−0.05	−0.04	0.04	0.11	0.01	−0.06	−0.02	−0.04
		(2.69)	(2.79)	(2.84)	(2.16)	(2.09)	(2.02)	(1.82)	(1.82)	(1.93)

Reported are the average differences of the log predictive scores of the three methods when compared against the random walk with drift. Values are shown for the variables employment (EMPL), consumer price index (CPI) and federal funds rate (FED) at forecast horizons of one month, one quarter, six months, and one year ahead. The methods differ in the way λ is chosen, optimize the fit in a presample (BGR), maximize the posterior likelihood (GLP), or fixed value $\lambda = 0.2$ (Sims). The results are obtained using different model sizes, Small: 3 variables, Medium: 7 variables, Large: 20 variables. All BVARs are estimated with five lags ($p = 5$) and predictive densities are evaluated based on 1000 draws. Positive values indicate the BVAR performed better than the benchmark. Values in parentheses are HAC standard errors.

significant patterns. Most of the scores are positive, which indicates that the three BVARs outperform the benchmark. However, not even the most positive results surpass half a standard error. Which does not support the notion that these differences are statistically significant. Overall, the results in Table 2 are further evidence that the three methods are very close in their forecasting ability, even when the whole predictive density is considered.

Next we will take a closer look at two specific periods where all methods perform very poorly, to see if the high forecast errors are the result of the models failing to capture the macroeconomic dynamics, or an exogenous shock, or if there is a different explanation. The first occasion is of two subsequent large forecasting errors, which seem to be the result of a sudden dip in employment in August 1983. Figure 1 shows a plot of the employment variable from 1982 to 1985 with corresponding forecasts made using the different methods. The figure shows that over the course of 1982 the variable is tough to forecast on account of the recession, this is accompanied by an irregular downward slope that is difficult stick close to. After the economy starts picking up in the first quarter of 1983 the prevailing

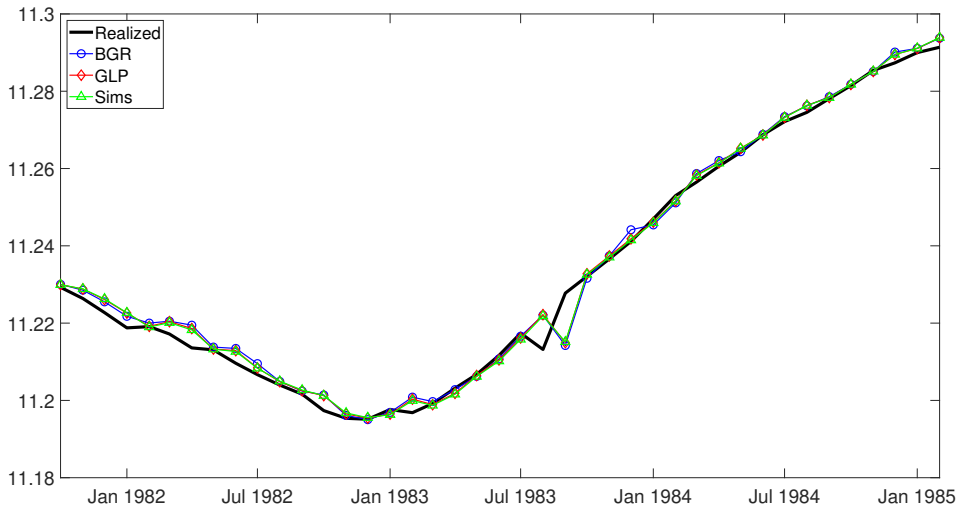
trend is in an upward direction and forecast errors are very minimal, until suddenly there is drop in August, which appears to be gone entirely the next month. Since the drop is so unexpected we immediately incur a large forecast error as the models overshoot their prediction. In the following period the models have incorporated the latest information and have adjusted accordingly, only to find that employment has reverted to a level that coincides with the trend that was established in the first half of 1983, resulting in another large forecast error. It is unlikely that this is the result of an exogenous shock, since it appears to affect only one observation, it is more plausible that this is caused by an observational error. Indeed, in *The Employment Situation* of August 1983, a monthly report that is published by the Bureau of Labor Statistics, we read that during this month 700,000 communications workers went on a three-week strike. The variable used in our analysis is defined as the total number of workers on private non-farm payrolls. Since workers on strike are not on any payroll, they were not counted for this observation, this resulted in a one-time observational error.

What we further see from this figure is that the different methods are very alike in their forecasts. Even though there are substantial differences in the amount of shrinkage between methods for this model, recall that the BGR imposes no shrinkage on the Small model, there is not much difference to be observed in the actual forecasted values. Another striking feature of this figure, is the disparity between the accuracy of the forecasts at the beginning of this subsample and at its end. The apparent cause of this is the change in economic climate that occurred around the start of 1983. We enter the subsample in the middle of a recession that started in 1981. Recessions are often accompanied by macroeconomic turbulence, which makes forecasting more difficult. When the economy starts picking up again, we see that employment is characterized by a clear upward trend, all three methods manage to stay close to it, and we observe very minimal forecast errors.

The second period of interest is the three years between 1980 and 1983. The federal funds rate during this time is characterized by extremely high volatility, which makes it difficult to model (see Figure 2). Since this rate is set by the Federal Reserve System, its dynamics are quite different from those of other variables. To understand the large fluctuations in the interest rate, one has to look at the rate of inflation during this time. Following the 1973-1975 recession, inflation remained high throughout the rest of the '70s. When the energy crisis hit in 1979, inflation shot up even further and the US economy went into a recession. One of the functions of the federal funds rate is to offset the effects of high inflation, because of this the FED often follows the movements of the rate of inflation. This explains why we observe such large changes that result in big forecast errors during this period.

Contrasting with Figure 1, there are definite differences across methods, in particular we see that the largest forecast errors are realized by the BGR model. Indeed, when

Figure 1: A dip in employment in August of 1983.



Plotted is a subsample showing the observations for log employment and the forecasts made by the different methods at one step ahead using the Small model.

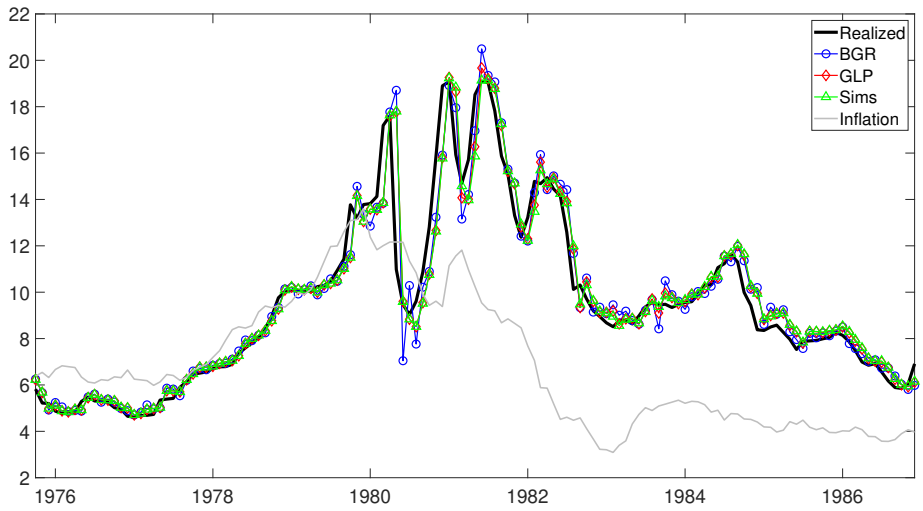
we revisit Table 1 and Table 2, we can see that they confirm that the BGR method is comparatively the worst in predicting with the Small model. For the federal funds rate it even fails to beat the random walk benchmark at any horizon. This may indicate that even in a BVAR with as few as three variables, forecasting performance benefits from any amount of shrinkage.

It was previously mentioned that changes in the FED are largely dictated by movements in the rate of inflation. To highlight this, Figure 2 also plots the year-on-year inflation during this period. If we direct our attention to the inflation series, another pattern emerges. During the first few years of this subsample, the FED is kept around the same level as inflation and the predicted values are very close to the observed values. Then we enter the period of high volatility that was discussed before, after which the inflation gradually decreases, but the interest rate does not follow suit. For the remainder of this subsample, the FED stays at a much higher level and we can see that this coincides with larger forecast errors. Even though both variables are at a stable level and changes are relatively small, as long as there is a discrepancy between the two variables forecasting performance remains poor.

5 Robustness analysis

In order to verify the robustness of the previous results, we now investigate the sensitivity of results to changes in two model parameters. Up until this point we have failed to give justification for our choice in the number of lags included in the VAR models. The decision

Figure 2: High volatility in the Federal Funds Rate during the early 80's.



Plotted is a subsample showing the observations for the Federal Funds Rate and the forecasts made by the different methods at one step ahead using the Small model. Also plotted is the year-on-year rate of inflation.

to use a lag order of five is mainly motivated by computational considerations. With a lag order of thirteen (the preferred choice in Bańbura et al., 2010) the number of parameters to estimate for the Large model is over five thousand. However, estimating the Small and Medium models with thirteen lags remains feasible, Table 3 shows the MSFE ratios relative to the benchmark.

Table 3: Robustness check for the number of lags, relative MSFE of point forecasts.

		Small			Medium		
Horizon	Variable	BGR	GLP	Sims	BGR	GLP	Sims
$h = 1$	EMPL	0.89	0.54	0.52	0.64	0.58	0.59
	CPI	0.97	0.50	0.50	0.55	0.51	0.52
	FED	1.42	0.91	0.93	0.85	0.86	0.84
$h = 3$	EMPL	0.73	0.48	0.45	0.60	0.53	0.55
	CPI	0.83	0.47	0.46	0.52	0.45	0.47
	FED	1.36	1.13	1.10	0.96	0.92	0.92
$h = 6$	EMPL	0.81	0.58	0.54	0.80	0.65	0.71
	CPI	0.89	0.51	0.49	0.52	0.45	0.48
	FED	1.82	1.36	1.28	1.19	1.02	1.05
$h = 12$	EMPL	0.75	0.65	0.62	1.02	0.81	0.88
	CPI	1.03	0.57	0.56	0.58	0.51	0.54
	FED	2.34	1.58	1.42	1.65	1.29	1.39

As a robustness check for the number of lags in the BVARs, the Small and Medium models are run with thirteen lags ($p = 13$). Values shown are the MSFE's relative to the MSFE of a random walk forecast. Results are reported for the variables employment (EMPL), consumer price index (CPI) and federal funds rate (FED) at forecast horizons of one month, one quarter, six months, and one year ahead. The methods differ in the way λ is chosen, optimize the fit in a presample (BGR), maximize the posterior likelihood (GLP), or fixed value $\lambda = 0.2$ (Sims).

When we compare these results to those presented earlier in Table 1 we see that there are very few differences. In most instances the forecasting performance relative to the random walk with drift barely changes. The notable exception to this is the first column with results for the Small model with BGR shrinkage. There we see that the additional lags have led to a decrease in forecasting ability for all three variables at all horizons. This can be explained through how the shrinkage is set. The method as it is defined in Bańbura et al. (2010) will always select $\lambda = \infty$ for the Small model, which means that in this case there is no shrinkage. With a model size of only three variables, parameter instability is not a large concern, unless the number of lags becomes too large. With thirteen lags, the number of coefficients to estimate has already grown to 120. The results in Table 3 indicate that at that point, forecasting performance suffers in the absence of shrinkage. For the other two methods we see that the MSFE ratios show signs of improvement in the majority of cases, however, these improvements are marginal.

The predictive likelihood scores for Table 2 have been calculated with 1000 draws from the predictive density. The choice for this value also follows from concerns over computation times, especially in the case of the Large model. To test the robustness of results to the number of draws, we now calculate the predictive scores for the Small model using 10,000 draws from the predictive density. These results are in Table 4.

Table 4: Robustness check for the number of draws from the predictive density, average difference of log predictive scores.

Horizon	Variable	BGR	GLP	Sims
$h = 1$	EMPL	0.30	0.30	0.29
		(0.70)	(0.65)	(0.62)
	CPI	0.30	0.37	0.36
		(1.01)	(0.89)	(0.89)
	FED	0.12	0.11	0.11
		(0.38)	(0.35)	(0.32)
$h = 3$	EMPL	0.43	0.35	0.35
		(1.44)	(1.45)	(1.42)
	CPI	0.56	0.65	0.65
		(2.20)	(2.04)	(2.02)
	FED	−0.10	−0.03	−0.01
		(1.00)	(0.96)	(0.94)
$h = 6$	EMPL	0.67	0.55	0.54
		(2.40)	(2.66)	(2.56)
	CPI	1.13	1.27	1.29
		(3.81)	(3.52)	(3.45)
	FED	−0.18	−0.12	−0.15
		(1.46)	(1.62)	(1.77)
$h = 12$	EMPL	1.43	1.22	1.09
		(4.43)	(4.78)	(4.74)
	CPI	3.29	3.75	3.78
		(8.79)	(8.95)	(8.88)
	FED	0.00	−0.03	0.02
		(2.69)	(2.79)	(2.82)

Reported are the average differences of the log predictive scores of the three methods when compared against the random walk with drift. Values are shown for the variables employment (EMPL), consumer price index (CPI) and federal funds rate (FED) at forecast horizons of one month, one quarter, six months, and one year ahead. The methods differ in the way λ is chosen, optimize the fit in a presample (BGR), maximize the posterior likelihood (GLP), or fixed value $\lambda = 0.2$ (Sims). The BVARs are estimated with five lags ($p = 5$) and the predictive densities are based on 10,000 draws. Results are shown only for the three-variable model, because of computational feasibility issues. Positive values indicate the model performed better than the benchmark. Values in parentheses are HAC standard errors.

Since there is no change in the model specification or the method of estimation, we expect there to be no differences between Table 2 and Table 4. This exercise merely serves to verify that 1000 draws is a large enough sample to not have to worry about sensitivity to the seed of the random number generator. Indeed, we find barely any differences between the two tables, neither in the log predictive scores, nor in the standard errors.

6 Posterior distributions

6.1 Posterior distribution of the hyperparameter

As mentioned before, the literature provides both empirical (Bańbura et al., 2010; Giannone et al., 2015) and theoretical (De Mol et al., 2008) evidence regarding the shrinkage of large Bayesian VARs. It has been shown conclusively that in order to protect against high

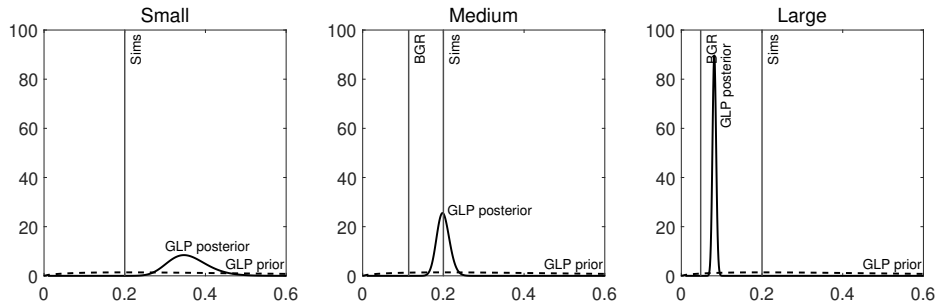
estimation uncertainty and overfitting, priors have to be put tighter as the dimensionality of the BVAR increases. That is why we now turn to investigate the differences in shrinkage between models and methods.

The values of the shrinkage parameter, λ are plotted in Figure 3. Since the BGR and Sims methods both select a fixed value at the start of the procedure, there is no difference in prior and posterior and all probability mass lies at one point. The hierarchical GLP method puts a proper hyperprior on λ and also produces a posterior distribution for the hyperparameter. The figure has a subplot for each model size. This makes it easy to see that as we consider larger BVARs, the methods that automatically infer the tightness opt for increasingly tighter priors. In the first figure, the value for BGR is not visible as it is by definition set to infinity, but in the succeeding plots we can see the selected values move closer to zero. Since the BGR values are chosen to maintain a certain level of fit this shows that more shrinkage is required to properly fit a larger model. The same is true for the GLP posterior distributions, while the imposed hyperprior remains the same across models, both the posterior mode and variance get smaller as the number of variables in the model increases. Lastly we note that the GLP hyperprior is sufficiently uninformative, it is largely flat over the range of reasonable values. As to differences between the methods it is interesting to note that for the Small model there are large differences in chosen values across methods. However, we saw before that the forecasting ability of BGR with the Small model is only slightly worse, indicating that the amount of shrinkage is not of considerable importance in a three-variable model. For the Medium model, both BGR and GLP arrive at values very close to Sims' fixed 0.2, this is attested by the great similarity in performances that we observed for the Medium model in the previous section. Lastly, for the Large model we see that both dynamic methods have moved away from the Sims value. Combined with what we noted on the last columns of Table 1 this proves that for BVARs of this size a higher degree of shrinkage significantly benefits the forecasting performance.

6.2 Posterior distribution of the VAR coefficients

In this section we look at the distributions of some of the VAR coefficients. Figure 4 and Figure 5 shows the distributions for the coefficients on two lags of EMPL, CPI, and FED, in the equations from the Medium model corresponding to these same variables. Plotted in these figures are the prior and posterior distributions of both the BGR and GLP methods. The previous section showed that the GLP and Sims method select the same amount of shrinkage for the Medium model and as a result their coefficient posteriors look almost identical, to avoid overcrowding in these figures we present only plots for BGR and GLP. From the previous section we know that BGR selects tighter priors, this is also apparent from these figures. Especially for the fourth lag the BGR priors can be noted to have

Figure 3: Posterior distribution of the shrinkage parameter of the Minnesota prior.



The figures shows plots of the posterior distribution of the shrinkage parameter λ for the BVARs of different sizes as determined by the hierarchical prior framework. The presented results are for the entire sample period 1960 - 2003. Also included is the hyperprior, a Gamma distribution with mode 0.2 and standard deviation 0.4. The values selected by the BGR and Sims methods are plotted as vertical lines, since they use fixed values. The BGR value for the Small model is $\lambda = \infty$, and is therefore not visible in the figure.

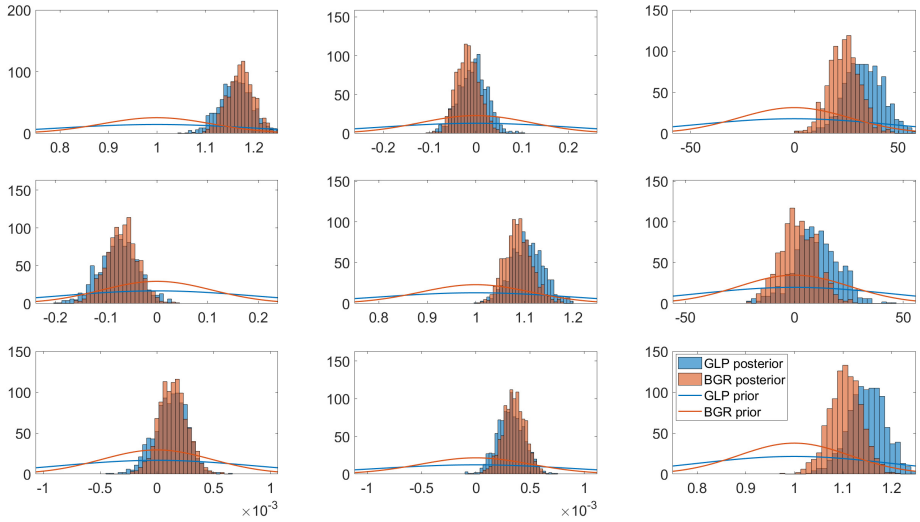
higher peaks than the GLP distributions. The posteriors are plotted as histograms of the samples $\{\boldsymbol{B}^{(i)}\}_{i=1}^{N}$ that were drawn when making draws from the predictive density. From the figures we can observe that the tighter prior of BGR results in narrower posteriors for the fourth lag coefficients, but for the first lags the BGR and GLP histograms seem to be of very similar shape. Most of the time there are slight differences in the mean value, but sometimes the two histograms almost completely overlap. The figures in the right column correspond to coefficients from the FED equation, there we see that there are significant shifts in location from prior to posterior distributions. This is also where the effect of the tighter BGR prior is most apparent as the mean manages to shift further away under the looser GLP prior. This may also indicate that the prior values for the mean are too far of their true values and need to be adjusted, or the priors on these coefficients should be loosened further.

When we make a comparison between the two figures we notice that the distributions for coefficients on fourth lags are a lot narrower. We recall that tightness of the priors increased quadratically with lag length, so the variance of the fourth lag priors is sixteen times as small. We see that in the first figure most posteriors move away from the mean value proposed by the prior. In the second figure a lot of the histograms are still centered around their zero prior values, but not all of them. This indicates that for most of the higher order coefficients the zero mean imposed by the prior is appropriate.

7 Structural analysis

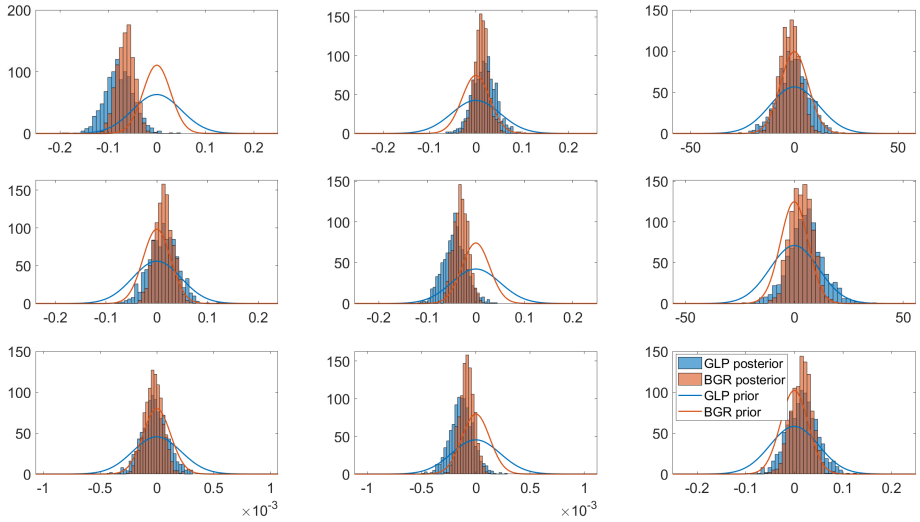
VARs are used in the literature not only for forecasting but also as a tool to identify structural shocks and assess their transmission mechanism. To this end we now investigate the IRFs, we do this by simulating an exogenous shock to the federal funds rate and plotting how the estimated models respond to this shock. The exercise consists of adding

Figure 4: Distributions of the coefficients on the first lags in the Medium model.



These figures show the prior and posterior distributions of selected VAR coefficients. The first row corresponds to the coefficients on the first lag of EMPL, the second and third row are for CPI and FED, respectively. The figures in the first column are for coefficients in the EMPL equation, the second and third correspond to the equations of CPI and FED, respectively.

Figure 5: Distributions of the coefficients on the fourth lags in the Medium model.

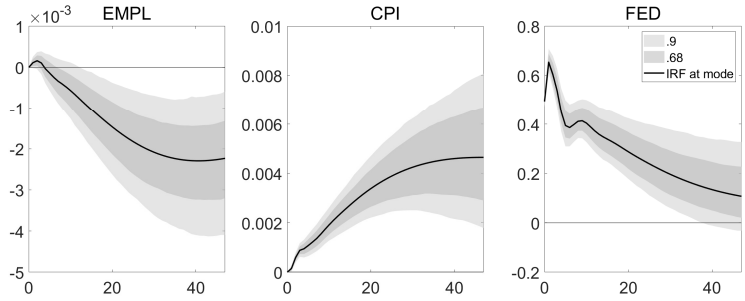


These figures show the prior and posterior distributions of selected VAR coefficients. The first row corresponds to the coefficients on the fourth lag of EMPL, the second and third row are for CPI and FED, respectively. The figures in the first column are for coefficients in the EMPL equation, the second and third correspond to the equations of CPI and FED, respectively.

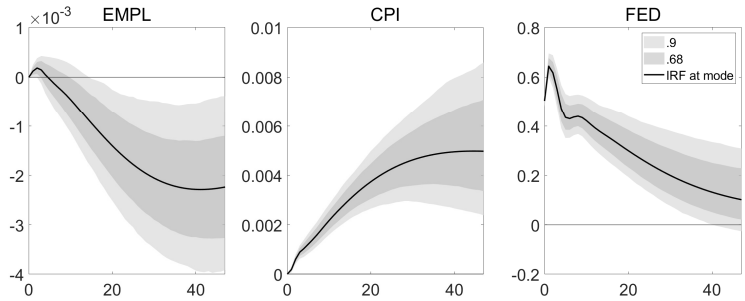
an exogenous shock of one standard deviation, which amounts to approximately 50 basis points, and calculating the effects over the next 48 months. Figure 6 shows the IRFs as calculated by the BGR method and the GLP method for the Small model using the entire sample. Plotted are the IRFs at the posterior mode, as well as the 0.68 and 0.90 coverage intervals. This figure mainly serves to demonstrate how similar the IRFs of the different methods are, this is also why there is no figure for the IRFs from Sims. Differences between BGR and GLP are so small that the plots seem indistinguishable.

Figure 6: Impulse response functions for the Small model.

(a) BGR



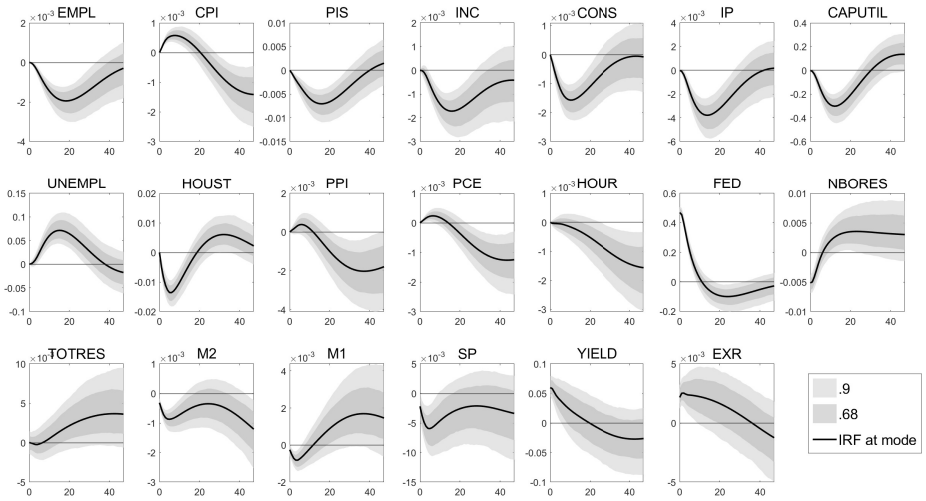
(b) GLP



These figures show the impulse response functions to a shock in monetary policy. This shock is created by an exogenous increase in the federal funds rate (FED) in the Small model. Plotted are the IRFs calculated at the posterior mode as well as the 0.68 and 0.9 coverage intervals, as estimated by the BGR and GLP methods.

Therefore, we move on to the discussion of the mechanisms as they are calculated with the Large model (Figure 7). The increase in the federal funds rate of half a percentage point simulates a contractionary change in monetary policy. As we would expect, the variables that measure real economic activity are negatively affected. Employment, income, consumption, industrial consumption and capacity utilization all show negative responses that persist for three years or more. An increase in the FED leads to overall higher interest rates, with the result that people spend less and businesses are less likely to expand now that borrowing has become more expensive. This causes the economy to slow down. We expect this reduction in economic activity to also lead to a higher unemployment rate, this expectation is confirmed by the graph for unemployment. Investors on the stock market are aware of the mechanisms of the economy and interpret an increase in the FED rate

Figure 7: Impulse response functions for the Large model calculated by the BGR method.



This figure shows the impulse response functions to a shock in monetary policy in the Large model estimated with BGR shrinkage. This shock is created by an exogenous increase in the federal funds rate (FED). Plotted are the IRFs calculated at the posterior mode as well as the 0.68 and 0.9 coverage intervals.

as a signal that business expansions will decline. The more pessimistic prospective view of the economy is quickly incorporated in stock prices. This is why we see an immediate drop in the S&P index. The effect on interest rates becomes apparent from the graph of the yield on 10-year treasury bond, the initial increase is large but fades rather quickly. The higher borrowing rates also has consequences for the housing market, as mortgages become more expensive the number of housing starts drops sharply. But after a year, when the interest rates move back to their previous levels, the housing market recovers.

The FED is a useful tool in combating inflation, when a rise in the rate leads to a slow-down of the economy, overall demand for goods goes down and prices fall. Indeed, the graphs show that price levels go down, after some delay we see large drops in the consumer price index, the producer price index and the PCE deflator, these lower levels are maintained at least until the end of the four years considered. The effect on the index of sensitive material prices (PIS) is more immediate, but also more transient. The effective exchange rate appreciates and it takes more than three years to revert to the pre-shock level. This is in line with the result from Eichenbaum and Evans (1995) which shows that a contractionary monetary policy shock leads to significant and persistent appreciation of exchange rates. Interesting to note is that the effect of the shock on the federal funds rate itself is highly transitory, while the FED is back at its original level after one year, the impact on other variables is often much more persistent.

Furthermore, when we compare the graphs produced with the Small model to the corresponding graphs in Figure 7, we notice that responses become more transient with the

additional information. This pattern is also observed in Bańbura et al. (2010), who note that when variables are added to the model the non-systematic component of monetary policy becomes smaller. To make the comparison easier, we include one final figure where the impulse response functions are plotted for the three key variables in the three models. If we take a row of subplots in Figure 8 and follow it from left to right, we can see how the shape of the IRF changes, in the cases of employment and the federal funds rate the patterns seems to be as described in Bańbura et al. (2010). The figures of the CPI are less straightforward. The Medium and Large figures correspond to previous results from the literature, but in the Small model the effect of the shock is calculated to be much larger than expected. Moreover, the effect is calculated to be persistent, the maximum horizon in these figures is four years but the response persists even beyond ten years. Finally, we note that the confidence intervals indicate that even though there is uncertainty in the size of the effect, which can be considerable in some cases, there is little doubt about its sign.

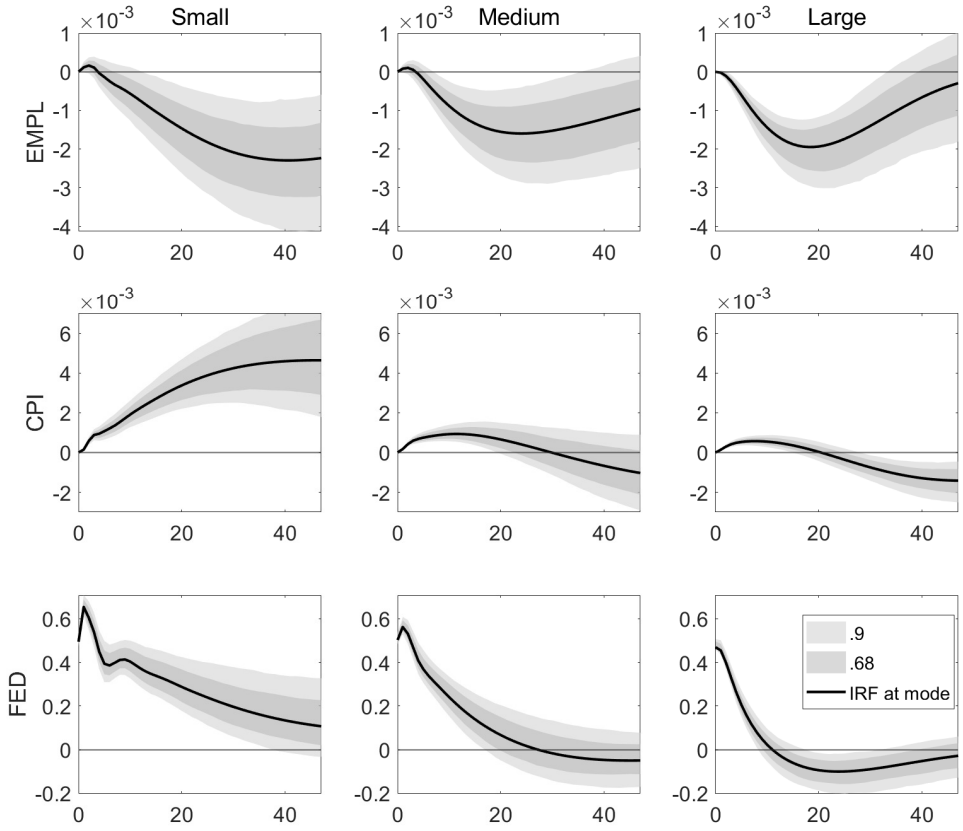
8 Conclusion

The goal of this research was to investigate whether their are any significant advantages to one of the two developed ways of setting the amount of shrinkage for a Bayesian vector autoregression. Both methods clearly choose to make the priors tighter as the model size increases, but there are differences in the exact amount of shrinkage. Results indicate that already in a small three-variable BVAR results can be improved by shrinking the model towards a random walk with drift. And for larger models considerable shrinkage is required to avoid the negative consequences of over-fitting and estimation uncertainty. But overall, the difference across methods in results such as forecasting performance, posterior distributions, or impulse response functions is minimal. Confirming Robert Litterman's conclusion that forecasting results are not overly sensitive to changes in the shrinkage parameter (Litterman, 1986).

Both methods have their own shortcomings, the method of Bańbura et al. (2010) as it is defined now does not impose shrinkage on the smallest model, even when moderate shrinkage might be appropriate. On the other hand, the hierarchical prior from Giannone et al. (2015) requires numerical optimization and the running of an MCMC algorithm at each iteration. Given that the methods are so close to each other in their results, the BGR method is much more useful when there are limitations to the available computing power.

One limitation of the application of the methods in this paper is the absence of additional priors that are often used in the VAR literature. The sum-of-coefficients prior and the dummy-initial-observation prior have both been shown to improve forecasting performance in BVARs. Adding these priors would introduce additional hyperparameters, in

Figure 8: Impulse response functions compared across models.



This figure shows the impulse response functions to a shock in monetary policy for the three different model sizes estimated with BGR shrinkage. This shock is created by an exogenous increase in the federal funds rate (FED). Plotted are the IRFs calculated at the posterior mode as well as the 0.68 and 0.9 coverage intervals.

order to keep this analysis as parsimonious as possible they have been omitted. However, if the goal were to compare the considered methods under optimal conditions, it would be appropriate to impose these additional priors. Even if there is little reason to believe that these extensions would qualitatively change the results, an extended analysis would provide a more complete answer to our question of interest.

References

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*(1), 71–92.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly journal of economics*, *120*(1), 387–422.

Canova, F. (1991). The sources of financial crisis: Pre-and post-Fed evidence. *International Economic Review*, 689–713.

Canova, F., & Ciccarelli, M. (2004). Forecasting and turning point predictions in a Bayesian panel VAR model. *Journal of Econometrics*, *120*(2), 327–359.

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? In J. B. Taylor & M. Woodford (Eds.), *Handbook of macroeconomics* (pp. 65–148). Elsevier.

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*(2), 318–328.

Dees, S., Mauro, F. d., Pesaran, M. H., & Smith, L. V. (2007). Exploring the international linkages of the euro area: A global VAR analysis. *Journal of Applied Econometrics*, *22*(1), 1–38.

Doan, T., Litterman, R., & Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, *3*(1), 1–100.

Eichenbaum, M., & Evans, C. L. (1995). Some empirical evidence on the effects of shocks to monetary policy on exchange rates. *The Quarterly Journal of Economics*, *110*(4), 975–1009.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, *82*(4), 540–554.

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian statistics*, *5*(599-608), 42.

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, *97*(2), 436–451.

Gordon, D. B., & Leeper, E. M. (1994). The dynamic impacts of monetary policy: An exercise in tentative identification. *Journal of Political economy*, *102*(6), 1228–1247.

Kadiyala, K. R., & Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, *12*(2), 99–132.

Litterman, R. B. (1979). *Techniques of forecasting using vector autoregressions* (tech. rep.). Federal Reserve Bank of Minneapolis.

Litterman, R. B. (1980). *A Bayesian procedure for forecasting with vector autoregressions*. MIT.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, *4*(1), 25–38.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48.

Sims, C. A., & Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, *39*(4), 949–968.

Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162.

Stock, J. H., & Watson, M. W. (2005a). An empirical comparison of methods for forecasting using many predictors. *Manuscript, Princeton University*, *46*.

Stock, J. H., & Watson, M. W. (2005b). *Implications of dynamic factor models for VAR analysis* (tech. rep.). National Bureau of Economic Research.

United States Bureau of Labor Statistics. (1983). *The employment situation: August 1983* (Monthly report on the employment situation USDL 83-387). United States Department of Labor. Retrieved April 23, 2021, from https://fraser.stlouisfed.org/title/employment-situation-144/august-1983-56163

Appendix

A Variables

Table 5: Descriptions of the variables in the dataset.

Mnemonic	Variable	Small	Medium	Log	RW	Slow/Fast
EMPL	Employees on non-farm payrolls	+	+	+	+	S
CPI	Consumer price index, all items less food and energy	+	+	+	+	S
FED	Federal funds rate	+	+		+	R
PIS	Index of sensitive material prices		+	+	+	S
NBORES	Non-borrowed reserves		+	+	+	F
TOTRES	Total reserves		+	+	+	F
M2	M2 Money stock		+	+	+	F
INC	Personal income less transfer payments			+	+	S
CONS	Real consumption			+	+	S
IP	Industrial production			+	+	S
CAPUTIL	Capacity utilization				+	S
UNEMPL	Unemployment rate				+	S
HOUST	Housing starts			+		S
PPI	Producer price index			+	+	S
PCE	Personal consumption expenditures price deflator			+	+	S
HOUR	Average hourly earnings			+	+	S
M1	M1 Monetary stock			+	+	F
SP	S&P common stock price index			+	+	F
YIELD	Yield on a 10 year US treasury bond				+	F
EXR	Effective exchange rate			+	+	F

This table gives a description of the dataset used throughout this research. The first two columns give the mnemonic through which the variable is identified and the description of the variable itself. The next two columns indicate whether a variable is included in the Small or Medium models. All variables are included in the Large model. The next column tells whether a log transformation is applied before entering the variable into the model, and the 'RW' column shows which variables were assigned the random walk prior. The last column tells whether a variable is considered slow or fast-moving for the purposes of the structural analysis: S is slow, F is fast, and R is the monetary policy instrument.

Table 6: Statistics for stationarity tests.

Variable	ADF	KPSS	PP
EMPL	−3.23 (0.08)	8.76* (0.01)	−0.79 (0.96)
CPI	−0.57 (0.98)	8.84* (0.01)	0.53 (1.00)
FED	−2.12 (0.53)	1.29* (0.01)	−2.26 (0.46)
PIS	−2.27 (0.46)	7.86* (0.01)	−1.35 (0.88)
NBORES	−1.08 (0.93)	8.47* (0.01)	−1.14 (0.92)
TOTRES	−0.32 (0.99)	8.51* (0.01)	−0.37 (0.99)
M2	−0.40 (0.99)	8.80* (0.01)	0.12 (1.00)
INC	−2.49 (0.35)	8.68* (0.01)	−1.96 (0.61)
CONS	−2.36 (0.41)	8.73* (0.01)	−2.03 (0.58)
IP	−2.98 (0.14)	8.33* (0.01)	−2.16 (0.51)
CAPUTIL	−3.57* (0.03)	1.51* (0.01)	−2.64 (0.28)
UNEMPL	−2.89 (0.17)	1.09* (0.01)	−1.92 (0.63)
HOUST	−3.49* (0.04)	0.21 (0.10)	−3.87* (0.01)
PPI	−0.40 (0.99)	8.59* (0.01)	0.29 (1.00)
PCE	−0.62 (0.98)	8.84* (0.01)	1.01 (1.00)
HOUR	0.11 (1.00)	8.71* (0.01)	1.33 (1.00)
M1	−0.82 (0.96)	8.91* (0.01)	−0.10 (0.99)
SP	−1.86 (0.66)	8.15* (0.01)	−1.66 (0.76)
YIELD	−1.64 (0.77)	1.97* (0.01)	−1.40 (0.86)
EXR	−2.26 (0.46)	4.81* (0.01)	−2.25 (0.47)

Test statistics of tests for stationarity with corresponding p-values in parentheses. Included are results for the augmented Dickey-Fuller test (ADF), the Kwiatkowski-Phillips-Schmidt-Shin test (KPSS), and the Phillips-Perron test (PP). The ADF test tests for the presence of a unit root in the sample, the alternative hypothesis is that of trend-stationarity. The null hypothesis of the performed KPSS test is that the time series is stationary, against the alternative of a unit root. Like ADF, PP tests for the presence of a unit root, the test statistic that is used is different, but the critical value is the same. All tests are run on models with five lags.

*: Null hypothesis is rejected at a confidence level of 5%.

B Derivation of posterior parameters

Consider the VAR from section 2 as written in (3)

$$Y = XB + U. \tag{B.1}$$

Since $U = (u_1, \ldots, u_T)'$, and $u_t \overset{iid}{\sim} N(0, \Sigma)$, we have $U \sim N(0, \Sigma \otimes I)$. Therefore, the data likelihood can be written as

$$p(Y|B, \Sigma) = (2\pi)^{-\frac{nT}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[(Y - XB)'(Y - XB)\Sigma^{-1}\right]\right). \tag{B.2}$$

Where $\text{tr}(\cdot)$ is the trace function. To make upcoming derivations easier, we already split this likelihood into two components, a conditional normal for B given Σ and an inverse-Wishart for Σ. We do this using the decomposition rule $(Y - XB)'(Y - XB) = (Y - X\hat{B})'(Y - X\hat{B}) + (B - \hat{B})'X'X(B - \hat{B})$, where $\hat{B} = (X'X)^{-1}X'Y$ is the ordinary least squares estimate of B. We also drop the integration constant and focus on the kernel:

$$\begin{aligned} p(Y|B, \Sigma) \propto & |\Sigma|^{-\frac{k}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[(B - \hat{B})'X'X(B - \hat{B})\Sigma^{-1}\right]\right) \\ & |\Sigma|^{-\frac{T-k}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[(Y - X\hat{B})'(Y - X\hat{B})\Sigma^{-1}\right]\right). \end{aligned} \tag{B.3}$$

The two lines above are the kernels of a Normal-Inverse-Wishart distribution with parameters

$$B|\Sigma, Y \sim N\left(\hat{B}, \Sigma \otimes (X'X)^{-1}\right), \tag{B.4}$$

$$\Sigma|Y \sim IW\left(\hat{S}, T - k - n - 1\right), \tag{B.5}$$

where $\hat{S} = (Y - X\hat{B})'(Y - X\hat{B})$. The likelihood functions of the prior distributions as given in (4) are as follows

$$p(B|\Sigma) = (2\pi)^{-\frac{kn}{2}} |\Sigma|^{-\frac{k}{2}} |\Omega_0|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[(B - \beta_0)'\Omega_0^{-1}(B - \beta_0)\Sigma^{-1}\right]\right) \tag{B.6}$$

$$p(\Sigma) = 2^{-\frac{nd_0}{2}} \frac{1}{\Gamma_n\left(\frac{d_0}{2}\right)} |S_0|^{\frac{d_0}{2}} |\Sigma|^{-\frac{d_0+n+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[S_0\Sigma^{-1}\right]\right) \tag{B.7}$$

Here $\Gamma_p(\cdot)$ denotes the multivariate gamma function. Multiplying the prior distributions in (B.6) and (B.7) with the data likelihood (B.3) we get a posterior likelihood with the

following kernel

$$p(B,\Sigma|Y) \propto |\Sigma|^{-\frac{k}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[(B-\hat{B})'X'X(B-\hat{B})\Sigma^{-1}\right]\right)|\Sigma|^{-\frac{T-k}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[\hat{S}\Sigma^{-1}\right]\right) \quad \text{(B.8)}$$
$$|\Sigma|^{-\frac{k}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[(B-\beta_0)'\Omega_0^{-1}(B-\beta_0)\Sigma^{-1}\right]\right)|\Sigma|^{-\frac{d_0+n+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[S_0\Sigma^{-1}\right]\right).$$

As a first step to finding the posterior parameters we gather the terms from the joint posterior that contain B. Then rearrange the terms so that coefficients can be matched

$$\exp\left(-\frac{1}{2}\text{tr}\left[\left((B-\hat{B})'X'X(B-\hat{B}) + (B-\beta_0)'\Omega_0^{-1}(B-\beta_0)\right)\Sigma^{-1}\right]\right) \quad \text{(B.9)}$$
$$= \exp\left(-\frac{1}{2}\text{tr}\left[\left((B'X'XB - 2B'X'X\hat{B} + \hat{B}'X'X\hat{B}) + (B'\Omega_0^{-1}B - 2B'\Omega_0^{-1}\beta_0 + \beta_0'\Omega_0^{-1}\beta_0)\right)\Sigma^{-1}\right]\right) \quad \text{(B.10)}$$
$$= \exp\left(-\frac{1}{2}\text{tr}\left[\left(B'(X'X+\Omega_0^{-1})B - 2B'(X'Y+\Omega_0^{-1}\beta_0) + \hat{B}'X'X\hat{B} + \beta_0'\Omega_0^{-1}\beta_0\right)\Sigma^{-1}\right]\right). \quad \text{(B.11)}$$

Where we used $\hat{B} = (X'X)^{-1}X'Y$ to get (B.11).

The posterior density function satisfies the same functional form, because the prior is a natural conjugate. So for the posterior we know we have a Normal-Inverse-Wishart distribution as well:

$$B|\Sigma, Y \sim N(\beta_1, \Sigma \otimes \Omega_1), \quad \text{(B.12)}$$
$$\Sigma|Y \sim IW(S_1, d_1), \quad \text{(B.13)}$$

with joint probability density function

$$p(B,\Sigma|Y) \propto |\Sigma|^{-\frac{k}{2}} \exp\left(-\frac{1}{2}\text{tr}[(B-\beta_1)'\Omega_1^{-1}(B-\beta_1)\Sigma^{-1}]\right)|\Sigma|^{-\frac{d_1+n+1}{2}} \exp\left(-\frac{1}{2}\text{tr}(S_1\Sigma^{-1})\right). \quad \text{(B.14)}$$

The goal is to match the coefficients of the normal density, the relevant part for this is the exponent term associated with the conditional normal kernel

$$\exp\left(-\frac{1}{2}\text{tr}[(B'\Omega_1^{-1}B - 2B'\Omega_1^{-1}\beta_1 + \beta_1\Omega_1^{-1}\beta_1)\Sigma^{-1}]\right). \quad \text{(B.15)}$$

Now the posterior parameters of the conditional normal can be found by matching coefficients between (B.11) and (B.15). By matching the quadratic term we find the posterior variance $\Omega_1 = (X'X + \Omega_0^{-1})^{-1}$. If we then match the terms that are linear in B and use the result of the posterior variance it turns out that $\beta_1 = (X'X + \Omega_0^{-1})^{-1}(X'Y + \Omega_0^{-1}\beta_0)$.

Once the terms associated with the kernel of the normal distribution are taken out of

the likelihood, this is what is left

$$p(\Sigma|Y) \propto |\Sigma|^{-\frac{T+d_0+n+1}{2}} \exp\left(-\frac{1}{2}\text{tr}[(S_0 + \hat{S} + \tilde{S})\Sigma^{-1}]\right), \tag{B.16}$$

$$\text{with} \quad \tilde{S} = \hat{B}'X'X\hat{B} + \beta_0'\Omega_0^{-1}\beta_0 - \beta_1'\Omega_1^{-1}\beta_1. \tag{B.17}$$

The first term is the product of all remaining $|\Sigma|$ terms, the second term gathers everything within the exponent, this includes the scale matrix of the prior inverse-Wishart distribution, the scale matrix from the data likelihood inverse-Wishart distribution, and a term that is the result of a discrepancy in the constant component after matching coefficients, $\tilde{S}$. Thus, we have found the parameters of the posterior inverse-Wishart distribution, scale matrix $S_1 = S_0 + \hat{S} + \tilde{S}$ and degrees of freedom $d_1 = T + d_0$.

C Additional Results

Table 7: Diebold-Mariano test statistics of the forecast errors.

		Small			Medium			Large		
Horizon	Variable	BGR	GLP	Sims	BGR	GLP	Sims	BGR	GLP	Sims
$h = 1$	EMPL	−4.43*	−5.29*	−5.60*	−4.42*	−4.45*	−4.09*	−5.06*	−4.87*	−4.07*
	CPI	−3.35*	−4.51*	−4.50*	−5.16*	−5.10*	−5.00*	−5.71*	−5.31*	−4.84*
	FED	0.76	−0.80	−0.55	−2.27*	−2.05*	−2.15*	−2.21*	−2.26*	−1.54
$h = 3$	EMPL	−2.72*	−2.77*	−2.90*	−2.31*	−2.33*	−2.14*	−2.61*	−2.61*	−2.29*
	CPI	−2.02*	−2.33*	−2.34*	−2.92*	−2.88*	−2.81*	−3.24*	−3.02*	−2.82
	FED	2.92**	1.40	1.18	−0.48	−0.38	−0.56	0.04	−0.14	0.07
$h = 6$	EMPL	−2.20*	−1.91	−2.06*	−1.62	−1.66	−1.42	−1.82	−1.78	−1.46
	CPI	−1.47	−1.92	−1.97*	−2.44*	−2.43*	−2.35*	−2.74*	−2.58*	−2.46*
	FED	1.79	1.28	1.12	0.16	0.17	0.13	1.30	1.24	1.68
$h = 12$	EMPL	−1.91	−1.19	−1.37	−1.12	−1.09	−0.89	−1.32	−1.01	0.12
	CPI	−1.09	−1.55	−1.56	−1.98*	−2.00*	−1.93	−2.45*	−2.37*	−2.20*
	FED	1.32	1.09	0.91	0.86	0.87	0.86	1.81	1.90	2.22**

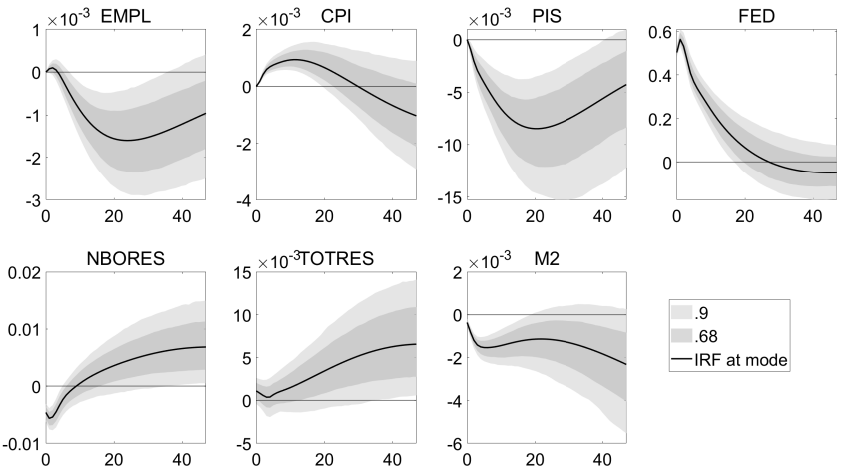
Presented are the Diebold-Mariano test statistics related to the results presented in Table 1. To correspond with the MSFE scores, the test statistics are calculated with a quadratic loss differential. Results are reported for the variables employment (EMPL), consumer price index (CPI) and federal funds rate (FED) at forecast horizons of one month, one quarter, six months, and one year ahead. The methods differ in the way λ is chosen, optimize the fit in a presample (BGR), maximize the posterior likelihood (GLP), or fixed value $\lambda = 0.2$ (Sims). The results are obtained using different model sizes, Small: 3 variables, Medium: 7 variables, Large: 20 variables. All BVARs are estimated with five lags ($p = 5$). Asterisks indicate whether the difference in predictive ability is significant based on the Diebold-Mariano test.

*: Null hypothesis of equal predictive accuracy is rejected at a confidence level of 5%.

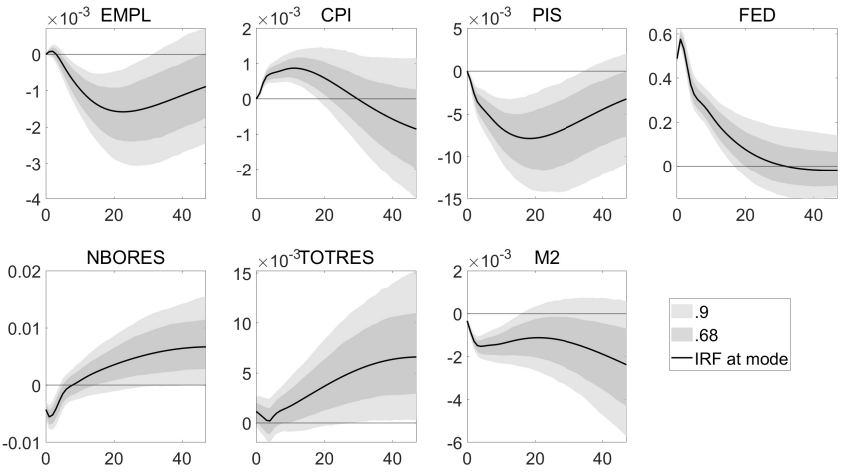
**: The null hypothesis is rejected in favour of the benchmark at a confidence level of 5%.

Figure 9: Impulse response functions for the Medium model.

(a) BGR



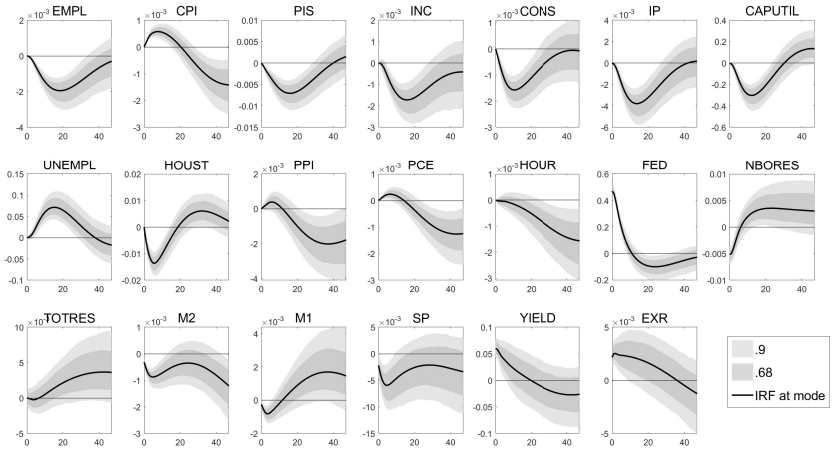
(b) GLP



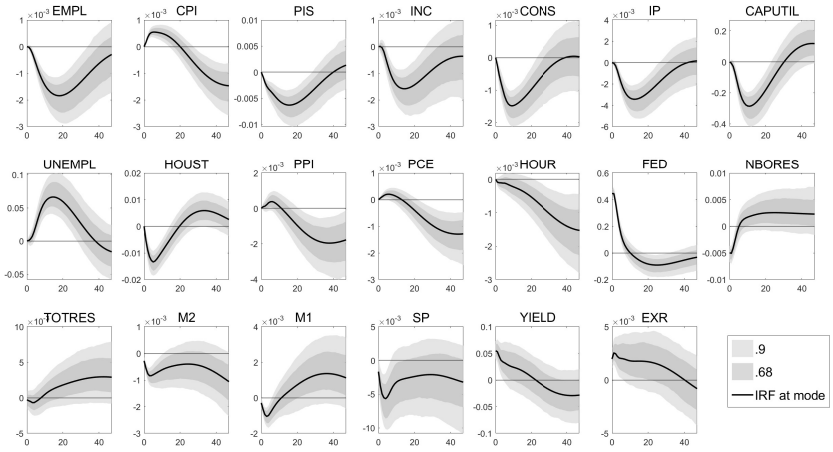
These figures show the impulse response functions to a shock in monetary policy. This shock is created by an exogenous increase in the federal funds rate (FED). Plotted are the IRFs calculated at the posterior mode as well as the median and the 16th and 84th quantiles.

Figure 10: Impulse response functions for the Large model.

(a) BGR

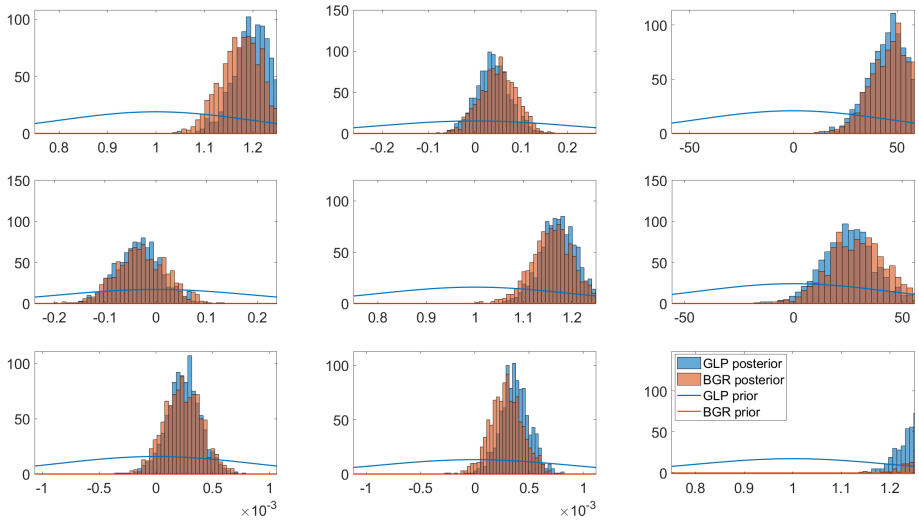


(b) GLP



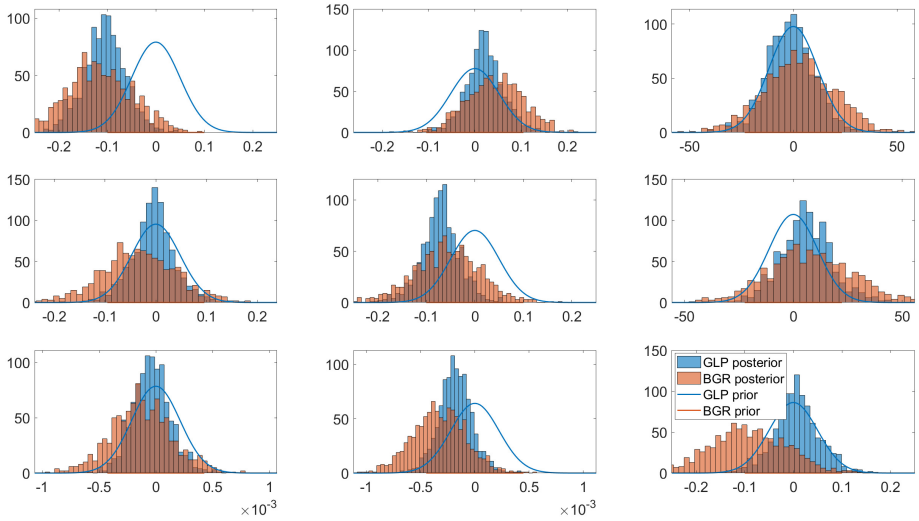
These figures show the impulse response functions to a shock in monetary policy. This shock is created by an exogenous increase in the federal funds rate (FED). Plotted are the IRFs calculated at the posterior mode as well as the median and the 16th and 84th quantiles.

Figure 11: Distributions of the coefficients on the first lags in the Small model.



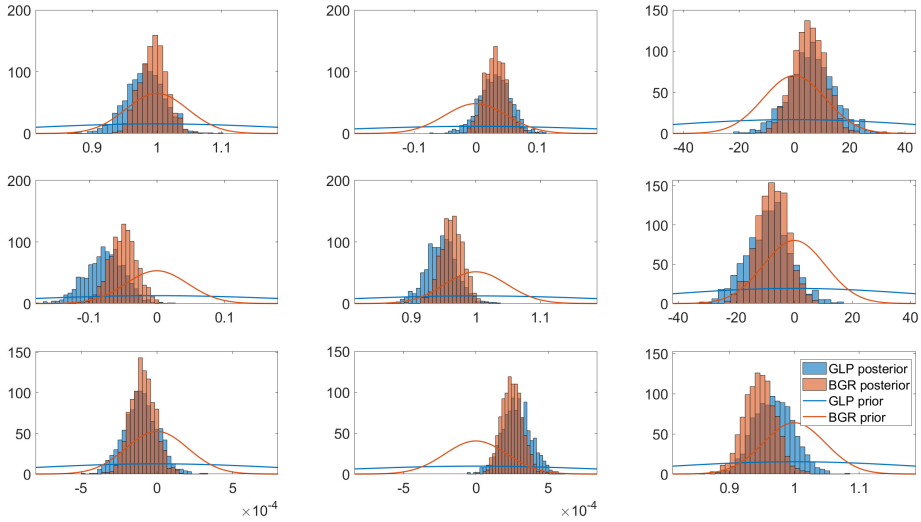
These figures show the prior and posterior distributions of selected VAR coefficients. The first row corresponds to the coefficients on the first lag of EMPL, the second and third row are for CPI and FED, respectively. The figures in the first column are for coefficients in the EMPL equation, the second and third correspond to the equations of CPI and FED, respectively.

Figure 12: Distributions of the coefficients on the fourth lags in the Small model.



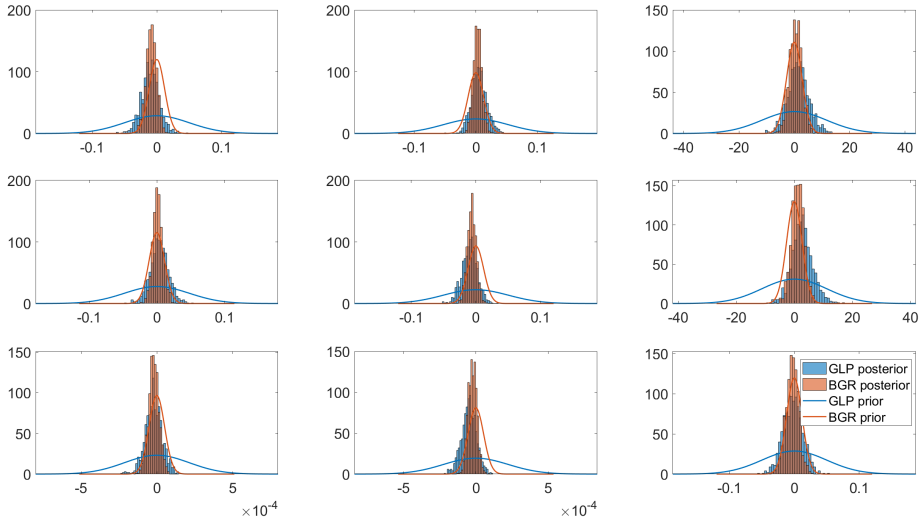
These figures show the prior and posterior distributions of selected VAR coefficients. The first row corresponds to the coefficients on the first lag of EMPL, the second and third row are for CPI and FED, respectively. The figures in the first column are for coefficients in the EMPL equation, the second and third correspond to the equations of CPI and FED, respectively.

Figure 13: Distributions of the coefficients on the first lags in the Large model.



These figures show the prior and posterior distributions of selected VAR coefficients. The first row corresponds to the coefficients on the first lag of EMPL, the second and third row are for CPI and FED, respectively. The figures in the first column are for coefficients in the EMPL equation, the second and third correspond to the equations of CPI and FED, respectively.

Figure 14: Distributions of the coefficients on the fourth lags in the Large model.



These figures show the prior and posterior distributions of selected VAR coefficients. The first row corresponds to the coefficients on the first lag of EMPL, the second and third row are for CPI and FED, respectively. The figures in the first column are for coefficients in the EMPL equation, the second and third correspond to the equations of CPI and FED, respectively.